JLSC

ISSN 2162-3309 | JLSC is published by the Pacific University Libraries | http://jlsc-pub.org

Volume 3, Issue 2 (2015)

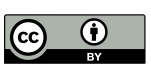
Do You Have an Institutional Data Policy? A Review of the Current Landscape of Library Data Services and Institutional Data Policies

Kristin Briney, Abigail Goben, Lisa Zilinski

Briney, K., Goben, A., & Zilinski, L. (2015). Do You Have an Institutional Data Policy? A Review of the Current Landscape of Library Data Services and Institutional Data Policies. *Journal of Librarianship and Scholarly Communication,* 3(2), eP1232. http://dx.doi.org/10.7710/2162-3309.1232

External Data or Supplements:

Briney, Kristin; Goben, Abigail; Zilinski, Lisa, 2015, "Data from: Do You Have an Institutional Data Policy? A Review of the Current Landscape of Library Data Services and Institutional Data Policies", http://dx.doi.org/10.7910/DVN/GAZPAJ, Harvard Dataverse



ISSN 2162-3309 10.7710/2162-3309.1232

RESEARCH

Do You Have an Institutional Data Policy? A Review of the Current Landscape of Library Data Services and Institutional Data Policies

Kristin Briney
Data Services Librarian, University of Wisconsin-Milwaukee
Abigail Goben
Information Services and Liaison Librarian, University of Illinois at Chicago
Lisa Zilinski
Data Services Librarian, Carnegie Mellon University

INTRODUCTION Many research institutions have developed research data services in their libraries, often in anticipation of or in response to funder policy. However, policies at the institution level are either not well known or nonexistent. **METHODS** This study reviewed library data services efforts and institutional data policies of 206 American universities, drawn from the July 2014 Carnegie list of universities with "Very High" or "High" research activity designation. Twenty-four different characteristics relating to university type, library data services, policy type, and policy contents were examined. **RESULTS** The study has uncovered findings surrounding library data services, institutional data policies, and content within the policies. **DISCUSSION** Overall, there is a general trend toward the development and implementation of data services within the university libraries. Interestingly, just under half of the universities examined had a policy of some sort that either specified or mentioned research data. Many of these were standalone data policies, while others were intellectual property policies that included research data. When data policies were discoverable, not behind a log in, they focused on the definition of research data, data ownership, data retention, and terms surrounding the separation of a researcher from the institution. **CONCLUSION** By becoming well versed on research data policies, librarians can provide support for researchers by navigating the policies at their institutions, facilitating the activities needed to comply with the requirements of research funders and publishers. This puts academic libraries in a unique position to provide insight and guidance in the development and revisions of institutional data policies.

Received: 02/27/2015 Accepted: 04/26/2015

Correspondence: Kristin Briney, Golda Meir Library, 2311 East Hartford Avenue, Milwaukee, WI 53211, briney@uwm.edu

IMPLICATIONS FOR PRACTICE

1. By leading the discussion on institutional and researcher needs, librarians can assist in creating policy as part of the overall data management infrastructure.
2. Understanding institutional policy will provide guidance in the development of research data services including data repositories and research collaborations.
3. Policy development can impact instruction through the application of data information literacy competencies based on needs and requirements of institution, funder, government, and publisher.

INTRODUCTION

Many research institutions' libraries have developed research data services and hired dedicated data librarians in order to collaborate with researchers to improve data management. However, while current published research data management curricula for librarians mentions data policy, the focus has primarily been on funder, government, and scholarly journal policies. As such, the place of library data services within the larger framework of research data support at the institutional level and in shaping emerging policies is often unclear.

Though major funding agency policies have received wide recognition, institutional level data policies are less well known, if they even exist. As a result, researchers may misunderstand their home institutional policies with regards to research data. This, in turn, leads to difficulties when the researcher wishes to or is required to share data; is attempting to coordinate policies in advance of a collaboration with researchers at another institution; or needs guidance setting policy for the department, unit, or lab.

The object of this study is to provide a comprehensive exploration of whether institutional research data polices can be located for larger United State (US) research institutions and determine any correlation between policy existence and either library data services or the presence of a data librarian. The results of this study will call attention to areas for future research in institutional data policies and the academic library's role.

LITERATURE REVIEW

Research in the 21st century is increasingly data-centric (Mullins, 2014). Policies surrounding research data are unclear, presenting obstacles for researchers to navigate. These policies, addressing ownership, retention, documentation, access, sharing, preservation,

and destruction, have great potential for overlap or contradiction, creating confusion. Many researchers are interested in reaping the benefits of data sharing and reuse, including training new researchers (Kriesberg, Frank, Faniel, & Yakel, 2013) and particularly in order to increase their impact and citation count (Piwowar, Day, & Fridsma, 2007; Tenopir et al., 2011). Beyond the time required and other challenges of preparing research data for archiving (Bruna, 2014), the lack of institutional policy or policy awareness creates confusion over best practices for multi-university collaboration (Jahnke, Asher, Keralis, Henry, & CLIR, 2012), appropriate attribution (Rohlfing & Poline, 2012), concern about ethical standards (Jahnke & Asher, 2013) and raises retraction fears (Marcus, 2013a, 2013b; Pittman, 2012) or other possibly litigable conflicts such as intellectual property disputes (Lewis, 2012; Schwartz, 2012).

Research data has become a topic of increased concern for librarians over the past 15 years and, in response, there has been an increase in resource and service development. Tenopir, Birch, & Allard (2012) provided the most recent comprehensive data services review in academic libraries in a white paper for the Association of College and Research Libraries (ACRL). The review identified that while few academic libraries offered data services in 2012, many were planning to launch them, with larger institutions leading the development of these services. Though the study identified a need for policies, including surrounding data retention, policy development was not identified as a major activity for librarians.

As these services have emerged, Antell, Foote, Turner, and Shults (2014) note library publications in this area have typically been case studies focused on librarians offering data services and assisting faculty with research management. Though more recent papers have mentioned the need for librarian training in policy, (Information School, University of Sheffield, Cox, Verbaan, & Sen, 2014; University of Manitoba & Ishida, 2014) they focus only on policies from funding agencies and other public policies without including institutional policies. Additionally, the role for librarians in creating and influencing policy is not explored in this research, leaving a gap in the opportunities for librarian influence. One of the popular data management curricula presently available for librarians, the New England Collaborative Data Management Curriculum (Sheridan et al., n.d.), includes a section about policies but the usefulness is limited in that it primarily focuses on sharing data and does not go into detail about reviewing institutional data concerns or clarifying ownership and retention at the institutional level.

Presently, there do not appear to be any reviews of challenges surrounding institutional data policies. In a recent book chapter, MacKenzie Smith (2014) outlines many of the general policy issues by examining what she calls data governance: lack of clarity in terms of licensing, inability to stack licenses to combine data sets, general practice in science

versus vague policy, different types of data, security and privacy for sensitive data (e.g. human subject), issues with creative works as data type, and software used in data creation and analysis. Though she briefly touches on ownership at the university level, her focus is primarily overarching themes such as terms of use, copyright and ownership licenses, public domain, and variances across international borders. According to Smith, these issues are impeding data sharing and reuse and institutions' abilities to create or improve tools in this area. Similarly, Krier and Strasser's book (2013) on data management for libraries only briefly covers institutional policies. Krier and Strasser recognize that many researchers start working at an institution without any knowledge of the institution's policies and that librarians must provide guidance in policy awareness and development. However, the brevity of the text does not provide sufficient direction for librarians looking to this avenue to effect change.

Research data ownership is particularly unclear and is frequently influenced by intellectual property laws, which are complicated by a variety of institutional, state, and country jurisdictions (Borgman, 2007; Jahnke & Asher, 2013). Some researchers assert that their data belong entirely to them (McGlynn, 2014), while others understand that their institution owns their data (Faulkes, 2014). However, many researchers are less clear about general data ownership rights (Murray-Rust, 2010). This can be further complicated with the variety of people who may be gathering data, including faculty, staff, research employees, students of all levels, fellow researchers, and collaborators at other institutions (Evola, 2013).

Law and government policy on the state, national, and international level heavily influence research data, especially in terms of access and intellectual property retention. In 2007, the *OECD Principles and Guidelines for Access to Research Data from Public Funding* (Pilat & Fukasaku, 2007) provided recommendations on access to data created with public funding for the 30 participating member governments, which included the United States. The Bayh-Dole Act (Law & Act, n.d.), an intellectual property law that frequently arises for the US, gives universities control of inventions funded by federal research money. Another frequently cited federal requirement is Circular A-100 from the Office of Management and Budget (1999), which was recently folded into the Office of Management and Budget's new Uniform Guidance (2014). This guidance states that the federal government has rights, including reproduction and publication, to data created from a federal grant. Finally, the Office of Science and Technology Policy released the "Increasing Access to the Results of Federally Funded Scientific Research" memorandum (Holdren, 2013), which directed US federal agencies to develop a plan to improve public access to federally funded research, including research data. During the spring of 2015, policies have begun to be released by the federal agencies (Adler, 2015; Whitmire et al., 2015). International research collaboration further complicates compliance as policies may or may not comprehensively exist.

Funding agency policies on research data are probably the best known by researchers due to the potential financial repercussions. One longstanding policy is the National Institutes of Health (NIH) Data Sharing Policy (8.2.3.1) (2010), which supports data sharing following acceptance of the primary research findings for publication. More recently, effective in 2011, the National Science Foundation (NSF) added to their data sharing policy (2010), which requires a data management plan where researchers outline how they will comply with the NSF sharing policy. Federal funding agencies have great variance in their policies (Dietrich, Adamus, Miner, & Steinhart, 2012) but a recent paper suggested that grant reviewers are not putting particular value on these data sharing plans (Pham-Kanter, Zinner, & Campbell, 2014). An analysis by Mischo, Schlemback, and O'Donnell (2014) revealed no significant difference in plans for funded and unfunded proposals. Various private funders have also adopted polices surrounding research data. One prominent example, which took effect on January 1, 2015, was the Bill & Melinda Gates Foundation (n.d.) requirement that all new funded agreements would immediately make published results and the supporting research data accessible to all. Further data policy developments from funders are expected in the coming years.

Scholarly journals have emerging and evolving policies regarding sharing and access to data that also affect researchers. A 2007 editorial ("Got data?," 2007) reminded authors that data sharing was part of the *Nature* author guidelines and recommended specific national repositories for consideration, though deposit is not mandatory for all dataset types (Nature Publishing Group, 2014). In 2014, PLOS took a stronger stance by requiring authors to make the data underlying their manuscripts available publicly or the article would be rejected (PLOS One, n.d.). This firm policy is similar to one long held by the American Geophysical Union for their accepted publications (American Geophysical Union, 2013). A more conservative example is the *Journal of the Medical Library Association*'s policy, which encourages data deposit and provides storage space for authors, but primarily focuses on retention (Medical Library Association, 2013). Many journals do not have an explicit data sharing policy. For further examples and reference points, Strasser provides short summaries about the Joint Data Archiving Policy (2012) and the Journal Research Data Policy Bank (2013).

Despite the many issues surrounding research data policy, there has been a limited amount of research focused on the content of those policies. Bohémier et al. (2011) surveyed 20 institutions with a Carnegie Classification of Very High to identify available data polices and their contents. However, the results included policies that were not institution wide and do not address ownership of data, and the small sample size limits the generalizability that can be extrapolated In 2013, researchers for the DataRes project (Keralis, Stark, Halbert, & Moen, 2013) briefly reviewed 197 institutions based on high NIH or NSF

funding to determine available policies. While this was a more comprehensive sample size, the researchers' returns were limited by use of only a few search terms to locate policies, and their analysis did not deeply explore the content or impact of the policies retrieved; instead, Keralis et al. summarize the policies as mostly ineffectual. The ARL SPEC Kit 334: Research Data Management Services (Fearon, Gunia, Lake, Pralle, & Sallans, 2013) provides a small selection of sample policies from major universities but presents older versions that did not yet respond to the emerging mandates.

Few papers provide context and framework, which could be drawn upon in creating or updating an institutional policy. Erway (2013) develops a detailed list of the stakeholders at institutions who need to be at the table when policies are being developed (e.g. IT, researchers, the Library, Office of Research), as well as identifying a variety of questions that need to be discussed over the course of policy development. Borgman (2012) reviews primary reasons for creation of data policy from the perspective of sharing. She also addresses an additional area: the reward motivation for researchers. Borgman points out that researchers are most likely to follow data research policies when there is a specific benefit to them. Furthermore, Keralis et al. (2013) point out institutions are likely to emulate funding agency policies, and the lack of momentum by the latter may be leading to lack of policy development by these institutions. Given the continually evolving nature of the data policy landscape, more work needs to be done to provide clarity and guidance in this area.

METHODS

The authors collected information on 206 American universities, their libraries' data services support efforts, and their institutional data policies. The list of universities chosen for this study was drawn from the July 2014 Carnegie list of universities in the United States with "Very High" or "High" research activity designation. Of the 206 universities, 107 were also identified as members of the Association of Research Libraries.

Student assistants gathered basic information on the 206 universities from the universities' websites, including type (public or private), total student population, faculty size, and research funding expenditures. This information was collected in June and July of 2014 and, wherever possible, data was from the 2012-2013 academic year. Two different student assistants collected data for each university to aid with accuracy. The authors cleaned the data by grouping faculty size and research funding into general categories to aid with analysis (Table 1, see Results). When there was disagreement as to which category a school should be placed into for a particular variable, the authors reexamined the university websites to decide proper categorization.

The authors gathered information from August to November of 2014 on which data support services the university libraries offered, including whether the library offered data services, had a data librarian on staff, maintained a repository specifically for data, or accepted data into their regular institutional repository (IR). Initially, all of the authors collected data on the same three universities to ensure consistency in collection. During this process, it was agreed that universities must offer services such as consultation or training to qualify as having data services (a data management LibGuide alone was not counted as data services); that a data librarian had data defined in her title or as a primary job duty; and that IR documentation must specifically mention data as an acceptable type of content for deposit. Beyond the initial three schools, data for the other 203 universities was collected by only one of the three authors.

Concurrent with collecting information on library data services, the authors searched for publicly available university data policies. Search phrases such as "data policy," "data retention," "data management," "data ownership," and "data stewardship" were used on university websites. Each university's intellectual property (IP) policy was also reviewed for explicit mention of "data." Policies that focused on institutional data (e.g. the common data set) or categorized levels of data requiring extra security such as student information were excluded, as these policies only applied to a subset of research data when it was covered at all. Similarly, data policies covering one department or college were not included. Policies that covered a whole university system were also excluded unless they were specifically hosted on an individual university's website. This method of identifying policies was chosen in order to estimate the ability of a researcher at the institution or an outside collaborator to locate a publicly-accessible policy. The authors did not attempt, for this project, to gain access to policies only available on intranets or behind other university log-ins. As with library services, all three authors coded the same three universities to ensure consistency in data gathering before having one researcher per university collect data. Gathered information on each policy included a link to the policy, a screenshot or saved pdf of the policy, and a record of which department or administrative office hosted the policy on the university website.
The final portion of data collection involved coding the policies for their content, specifically looking to see if the policy included any of the following:

- Policy defines research data in any way
- Policy states retention requirement of specific duration
    - What is that duration?
- Policy states retention requirement but does not specify duration
- Policy defines a data owner
    - Who is the data owner?

- Policy designates a data steward to administrate the data
- Policy states who is allowed/required to have access to the data
- Policy defines what happens to the data when the researcher leaves the university
- Data fall under the university's regular IP policy

As with previous data collection, inter-coder reliability was verified with a sample of the data before individual authors coded a subset of the universities. Coding was arranged so that the researcher who initially collected the policy information on the university website did not code the policy contents.

The authors collected all of the classification and coding data into a master spreadsheet covering all 206 universities. OpenRefine was then used to standardize the data and determine the final counts presented in the results section. Variables were counted individually (e.g. the total number of universities with library data services) as well as in combination (e.g. the number of universities with data services that also have a data policy). The authors performed a chi squared test for all variable correlations, using a threshold of $p = 0.05$ as a measure of significance.

RESULTS

The authors examined 206 universities by 24 different characteristics relating to university type, library data services, policy type, and policy contents. To best present this information, correlations are shown for the following sets of characteristics: university type and data services, university type and policy type, data services and policy type, and policy type and policy contents. These correlations correspond to the four subsections of the Results section. Each subsection contains tables of university counts and corresponding percentages for a set of characteristics with respect to another characteristic defined at the beginning of each row.

Library Data Services by University Type

Half (50%) of all the universities studied offered some data services through their library, with larger, more research-focused universities being more likely to offer data services than their counterparts. These findings are shown in Figures 1 and 2 (page 11). In particular, universities with a higher Carnegie classification ($\chi^2 = 63.13$, $p < 0.01$), ARL membership ($\chi^2 = 28.03$, $p < 0.01$), higher research expenditures ($\chi^2 = 58.91$, $p < 0.01$), and larger faculty size ($\chi^2 = 23.65$, $p < 0.01$) (as broken down in Table 1, following page) are all more likely to offer data services through their university libraries. Tables 2 and 3 (following pages) break down these findings in more detail. There is no significant difference between the percent of public and private universities that offer data services ($\chi^2 = 0.96$, $p = 0.33$).

Research Expenditures Categories	Faculty Size Categories
$0-25 Million	0-500
$26-100 Million	501-1000
$101-200 Million	1001-2000
$201-500 Million	Over 2000
$501-1000 Million	
Over $1 Billion	

Table 1. Data Coding Categories

Type	Total	Data Services	%	Data Librarian	%	Data Repository	%	Institutional Repository	%	Repository (Either Type)	%
Carnegie "Very High"	108	83	77%	65	60%	15	14%	71	66%	80	74%
Carnegie "High"	98	21	21%	12	12%	8	8%	48	49%	53	54%
ARL Member	107	73	68%	59	55%	15	14%	66	62%	77	72%
Not ARL Member	99	31	31%	18	18%	8	8%	53	54%	56	57%
Public Institution	145	70	48%	55	38%	19	13%	85	59%	97	67%
Private Institution	61	34	56%	22	36%	4	7%	34	56%	36	59%
Average	206	104	50%	77	37%	23	11%	119	58%	133	65%

Table 2. Library Data Services by University Type

Type	Total	Data Services	%	Data Librarian	%	Data Repository	%	Institutional Repository	%	Repository (Either Type)	%
$0-25 Million Research	21	4	19%	2	10%	1	5%	9	43%	9	43%
$26-100 Million Research	60	13	22%	8	13%	4	7%	31	52%	34	57%
$101-200 Million Research	42	23	55%	13	31%	5	12%	21	50%	23	55%
$201-500 Million Research	49	32	65%	27	55%	6	12%	33	67%	37	76%
$501-1000 Million Research	25	23	92%	19	76%	4	16%	18	72%	21	84%
Over $1 Billion Research	9	9	100%	8	89%	3	33%	7	78%	9	100%
0-500 Faculty	19	4	21%	1	5%	0	0%	8	42%	8	42%
501-1000 Faculty	52	19	37%	11	21%	8	15%	24	46%	29	56%
1001-2000 Faculty	77	38	49%	28	36%	7	9%	49	64%	53	69%
Over 2000 Faculty	58	43	74%	37	64%	8	14%	38	66%	43	74%
Average	206	104	50%	77	37%	23	11%	119	58%	133	65%

Table 3. Library Data Services by University Size

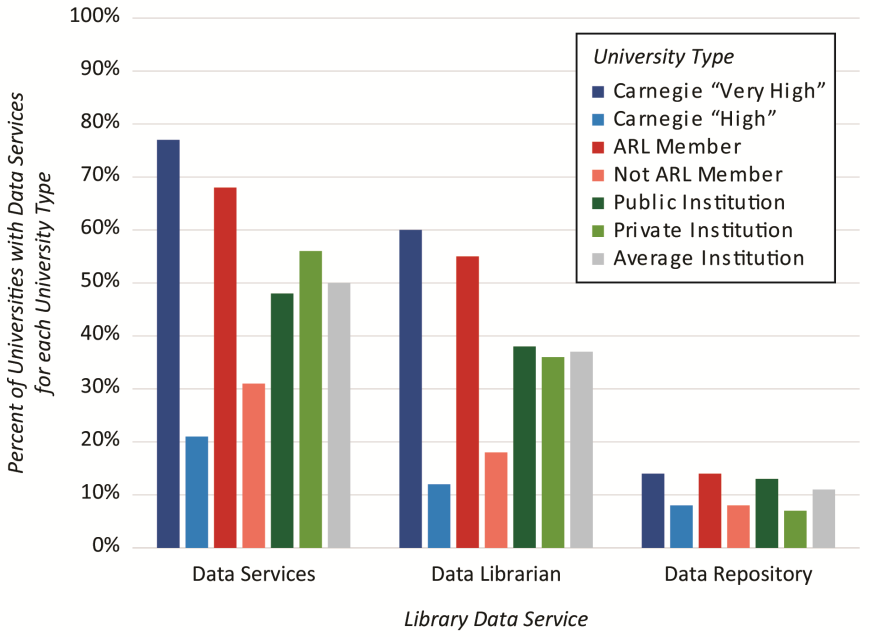


Figure 1. Existence of Library Data Services in Different Types of Universities

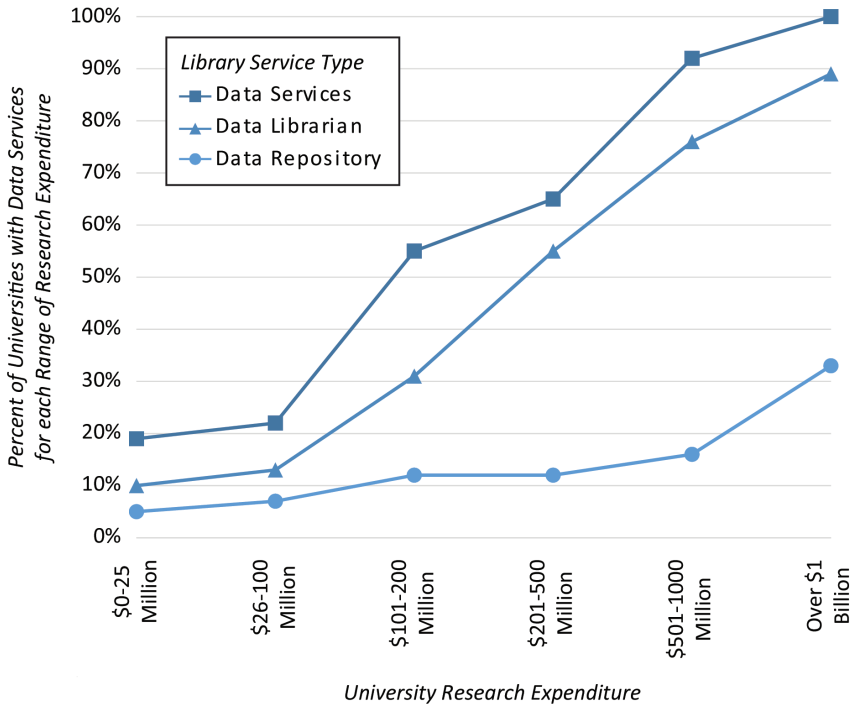


Figure 2. Existence of Library Data Services Broken Down by University Research Expenditure

Fewer universities have a data librarian on staff (37%) than offer data services. As with data services, the tendency is for larger universities conducting more research to employ a data librarian. This is true for universities with a higher Carnegie classification ($\chi^2 = 50.45$, $p < 0.01$), ARL membership ($\chi^2 = 30.01$, $p < 0.01$), higher research expenditures ($\chi^2 = 55.23$, $p < 0.01$), and larger faculty size ($\chi^2 = 31.54$, $p < 0.01$). Again, there is no significant difference between public and private universities ($\chi^2 = 0.06$, $p < 0.80$).

More than half of all libraries (65%) offer a place to host research data, either in an IR or in a repository specifically for data. However, fewer universities (11%) have dedicated data repositories as compared with IRs that accept data (58%). As with having data services and a data librarian, larger universities conducting more research are more likely to have someplace to host research data, as measured by Carnegie Classification ($\chi^2 = 8.98$, $p < 0.01$), ARL membership ($\chi^2 = 5.33$, $p = 0.02$), research expenditures ($\chi^2 = 19.36$, $p < 0.01$), and number of faculty ($\chi^2 = 8.88$, $p = 0.03$). For universities that accept data into their institutional repository, only Carnegie Classification ($\chi^2 = 5.92$, $p = 0.02$) and faculty size ($\chi^2 = 8.88$, $p = 0.03$) correspond having a higher likelihood of hosting data; ARL membership ($\chi^2 = 1.40$, $p = 0.24$) and research expenditure ($\chi^2 = 9.26$, $p = 0.10$) are non-significant. There are no significant correlations for universities with data repositories for Carnegie Classification ($\chi^2 = 1.70$, $p = 0.19$), ARL membership ($\chi^2 = 1.83$, $p = 0.18$), research expenditures ($\chi^2 = 8.24$, $p = 0.14$), or faculty size ($\chi^2 = 4.06$, $p = 0.26$).

Notably, all universities with over $1 billion per year in research expenditures offer data services and a place to host data. The vast majority (89%) of these institutions also have a data librarian on staff. Additionally, this group shows the highest chance of having a data repository (33%).

Data Policy by University Type

Out of 206 universities, only 90 had some type of university-level policy covering research data (44%). One-third of these university policies are IP policies that specifically include data (15% overall) while the remaining two-thirds are standalone data policies (29% overall). One university in particular, John Hopkins, has both an independent data policy and an IP policy that covers research data. Due to having both types of policies, John Hopkins' data is included in both the IP policy and a data policy columns of Tables 4 and 5 (following page) for all of the categories to which the school belongs (Carnegie "Very High", ARL member, private university, over $1 billion in research expenditure, and between 501-1000 faculty).

The data show universities that conduct more research are more likely to have some type of data policy (see Tables 4 and 5). This correlation exists across Carnegie Classification ($\chi^2 = 6.15$, $p = 0.01$) and ARL membership ($\chi^2 = 5.38$, $p = 0.02$), but not research expenditure

Type	Total	Has Data Policy	%	Data Under IP Policy	%	Data Under Other Policy	%
Carnegie "Very High"	108	56	52%	15	14%	42	39%
Carnegie "High"	98	34	35%	15	15%	19	19%
ARL Member	107	55	51%	17	16%	39	36%
Not ARL Member	99	35	35%	13	13%	22	22%
Public Institution	145	64	44%	23	16%	41	28%
Private Institution	61	26	43%	7	11%	20	33%
Average	206	90	44%	30	15%	60	29%

Table 4. Data Policy Type by University Type

Type	Total	Has Data Policy	%	Data Under IP Policy	%	Data Under Other Policy	%
$0-25 Million Research	21	8	38%	5	24%	3	14%
$26-100 Million Research	60	20	33%	7	12%	13	22%
$101-200 Million Research	42	16	38%	3	7%	13	31%
$201-500 Million Research	49	26	53%	8	16%	18	37%
$501-1000 Million Research	25	14	56%	6	24%	8	32%
Over $1 Billion Research	9	6	67%	1	11%	6	67%
0-500 Faculty	19	8	42%	3	16%	5	26%
501-1000 Faculty	52	21	40%	8	15%	14	27%
1001-2000 Faculty	77	31	40%	8	10%	23	30%
Over 2000 Faculty	58	30	52%	11	19%	19	33%
Average	206	90	44%	30	15%	60	29%

Table 5. Data Policy Type by University Size

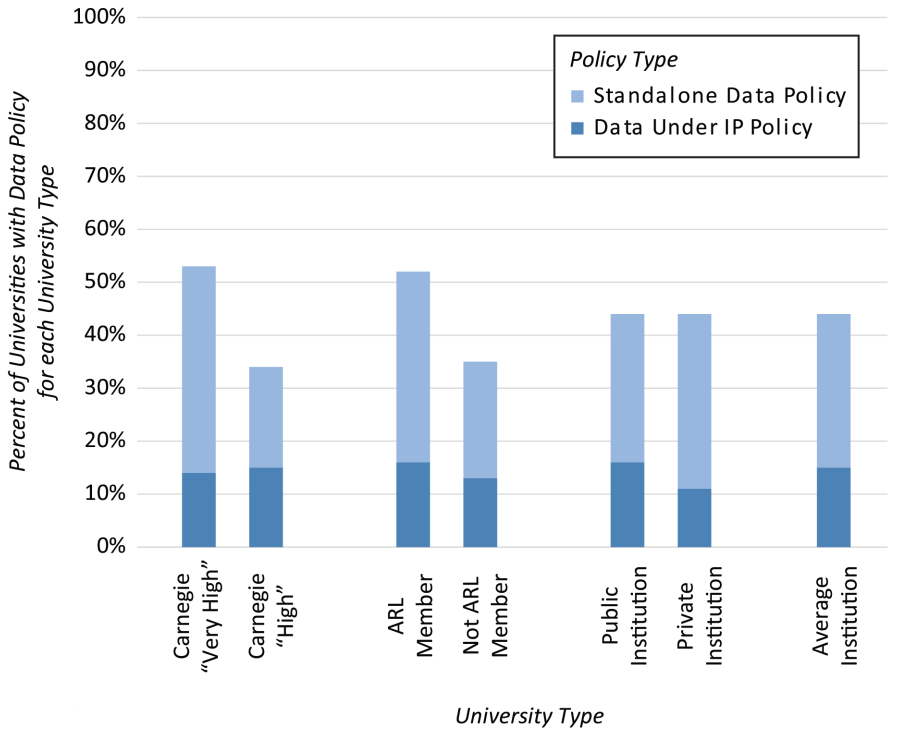


Figure 3. Existence of Data Policy in Different Types of Universities

($\chi^2 = 8.64$, $p = 0.12$) or faculty size ($\chi^2 = 2.14$, $p = 0.54$). There is no significant difference between public and private universities in this area ($\chi^2 = 0.04$, $p = 0.84$). In particular, the data show that universities conducting more research are more likely to have a standalone data policy. This is significant for Carnegie Classification ($\chi^2 = 9.37$, $p < 0.01$), ARL membership ($\chi^2 = 4.99$, $p = 0.03$), and research expenditure ($\chi^2 = 11.41$, $p = 0.04$). The same trends do not exist for university types when data falls under the normal IP policy; universities with higher Carnegie Classification ($\chi^2 = 0.08$, $p = 0.77$), ARL membership ($\chi^2 = 0.31$, $p = 0.58$), higher research expenditure ($\chi^2 = 5.70$, $p = 0.34$), and larger faculty size, ($\chi^2 = 2.03$, $p = 0.57$) show no significant difference from the average. Figure 3 visualizes these policy trends for Carnegie classification, ARL membership, and public/private universities.

Data Policies with Respect to Library Data Services

Nearly half (44%) of all universities studied have some type of policy covering research data, as reported in Table 6 (following page). Half of all libraries with data services have some data policy, but this is not a significant difference from the average ($\chi^2 = 3.40$, $p = 0.07$).

However, universities employing a data librarian are statistically more likely to have some type of data policy ($\chi^2 = 7.38$, $p < 0.01$). There is no significant difference in policy numbers for universities hosting data in a repository of any type.

Type	Total	Has Data Policy	%	Data Under IP Policy	%	Data Under Other Policy	%
Data Services	104	52	50%	16	15%	37	36%
Data Librarian	77	43	56%	14	18%	30	39%
Data Repository	23	7	30%	3	13%	5	22%
Institutional Repository	119	50	42%	15	13%	35	29%
Repository (Either Type)	133	56	42%	18	14%	39	29%
Average	206	90	44%	30	15%	60	29%

Table 6. Data Policy Type by Library Data Services

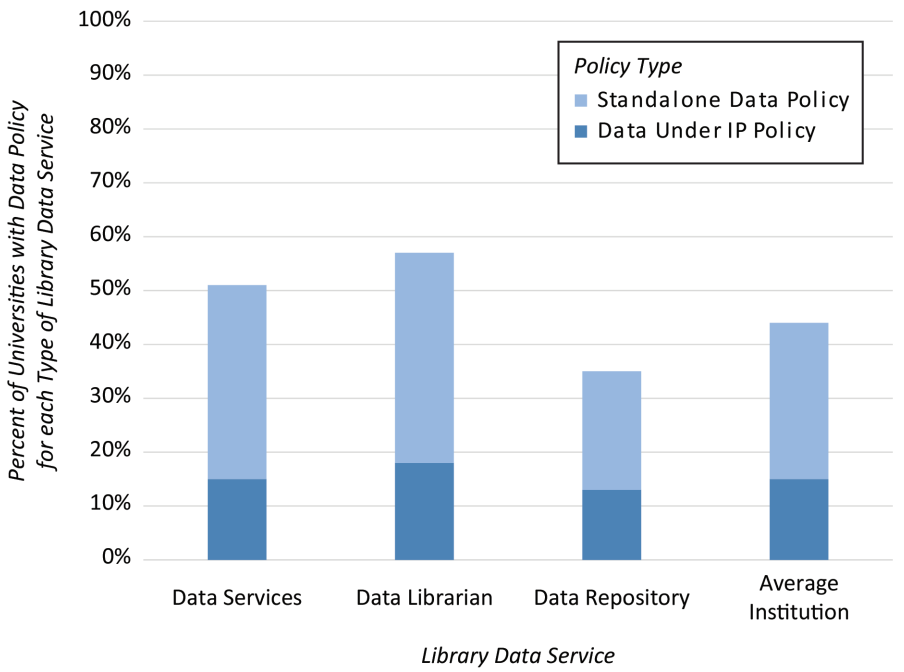


Figure 4. Existence of Data Policy for Universities with Library Data Services

Breaking the numbers down further by policy type, standalone data policies are more likely to be found at universities with data services ($\chi^2 = 4.23$, $p = 0.04$) and a data librarian ($\chi^2 = 5.76$, $p = 0.02$), but not those that host data in any repository ($\chi^2 = 0.01$, $p = 0.93$) or specifically a data repository ($\chi^2 = 0.68$, $p = 0.41$). Universities with data under their standard IP policy, however, show no significant difference in offering data services ($\chi^2 = 0.11$, $p = 0.74$), having a data librarian ($\chi^2 = 1.29$, $p = 0.26$), or hosting data in any repository ($\chi^2 = 0.32$, $p = 0.57$) or a data-specific repository ($\chi^2 = 0.05$, $p = 0.83$). Figure 4 shows the percent of universities with different data services that also have data policy, both standalone and IP policies covering data.

Again, as Johns Hopkins has both an IP policy covering data and a separate data policy, it is counted in both of these columns in Table 6 for all of the categories for which it belongs (data services, data librarian, and data repository).

Data Policy Contents

For the 90 universities that had data policies, the policy contents were coded by whether the policy defined data, identified a data owner, designated a non-owner responsible party (or data steward), required a retention period (either specific or vague), identified who is allowed access to the data, and described what happens to the data when a researcher leaves the university. Correlation statistics were not performed for these values as it was deemed outside of the scope of the current article. The authors plan to do a more comprehensive content analysis, including text analysis, in the future. Table 7 (following page) breaks down the contents of the policies by type of policy and Figure 5 (following page) displays the prevalence of the coded characteristics for different policy type. Note that in this case, the Johns Hopkins data is not included in the IP policy row as it was coded primarily by its standalone data policy.

Overall, over half of the policies designated an owner of research data generated at the university (67%) and required that the data should be retained for some period of time (43% for a specific period and 9% for a vague period).

There is a noticeable difference between the contents of IP policies that cover data and standalone data policies. IP policies primarily covered data ownership (76%) with little attention given to other data management issues. Standalone data policies, on the other hand, covered many topics. Over half of the data policies defined data (61%), identified a data owner (62%), state a specific retention time (62%), identified who can have access to the data (52%), and described what happens to the data when a researcher leaves the university (64%). Almost half of the policies (46%) also designate a data steward.

Type	Total	Defines Data	%	Identifies Data Owner	%	Appoints Data Steward	%
Has Policy	90	40	44%	60	67%	31	34%
Data Under IP Policy	29	3	10%	22	76%	3	10%
Data Under Other Policy	61	37	61%	38	62%	28	46%
Average	206	40	19%	60	29%	31	15%

Table 7a. Policy Contents by Data Policy Type

Type	Total	Requires Retention (Specific)	%	Requires Retention (Vague)	%	Defines Who Has Access to Data	%	Outlines Data Transfer If Researcher Leaves the University	%
Has Policy	90	39	43%	8	9%	34	38%	41	46%
Data Under IP Policy	29	1	3%	2	7%	2	7%	2	7%
Data Under Other Policy	61	38	62%	6	10%	32	52%	39	64%
Average	206	39	19%	8	4%	34	17%	41	20%

Table 7b. Policy Contents by Data Policy Type

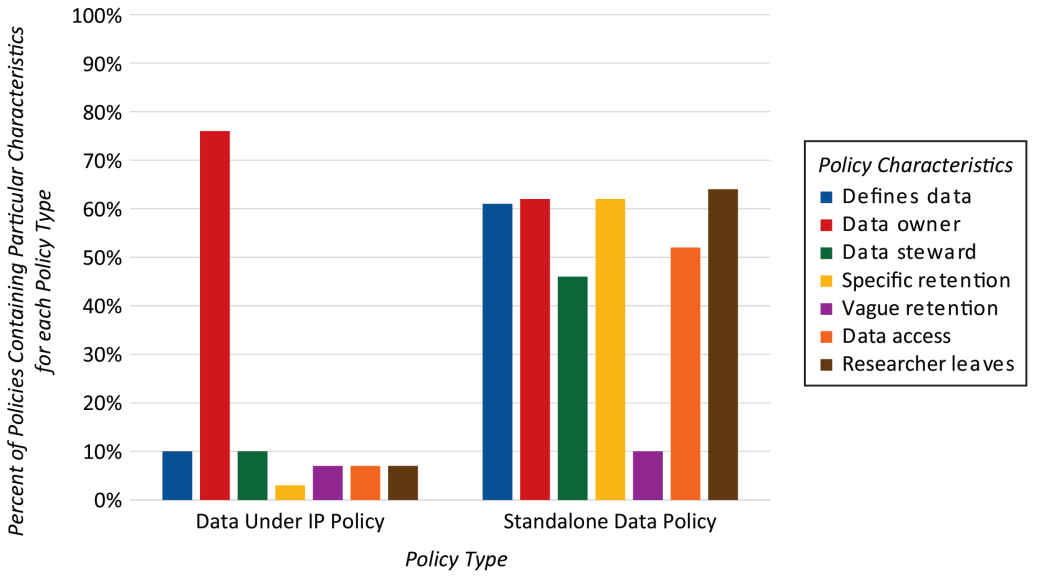


Figure 5. Policy Contents for Standalone Policies and IP Policies Covering Research Data

DISCUSSION

Library Data Services

The 2011 implementation of the NSF data management plan requirement was the impetus for a significant number of university libraries to create data services. Within only a few years of the requirement going into effect, half of the major research universities now offer data services. This is a large increase from the approximately 20% of ACRL libraries offering data-related services previously observed by Tenopir et al. (2012). However, this study's numbers are fairly consistent with their projection of research libraries that will offer data services in the near future. The comparison is not perfect, as the numbers in this study capture only overall data services while Tenopir et al. looked at individual services that may not be fully discoverable via the public websites for the libraries. Still, there is an overall increase in library data services being offered at research institutions. These results also confirm that the more research a university does the more likely its library is will offer data services. This demonstrates a general expectation for libraries to provide data services when their patrons conduct high levels of research.

While not as prevalent as data services, a large number of academic libraries have a data librarian on staff. This is again an increase over what was observed by Tenopir et al. (2012) who found that less than 10% of ACRL libraries had a dedicated data librarian/specialist. Duties for such librarians vary from offering data management support to assisting patrons with finding data for reuse, but the increasing number of data librarians is another sign that libraries are taking the new focus on data seriously.

It is somewhat surprising that more universities offer a place to host data than offer data services. An especially large portion of hosting comes from institutional repositories that accept some smaller data sets for deposit. Based on observations during data collection, the authors suspect that these large numbers are mainly due to hosted repository platforms, such as BePress, that state in their default guidance that the repository accepts data. While repository documentation states that data is accepted (and it was thus counted for this study), it is not fully clear if those libraries are actively acquiring data. Therefore, it is likely that the number of libraries actively working to collect research data is smaller than the reported 65%.

Unlike IRs, a small number of universities offer data repositories and/or promote them on their library webpages. While not statistically significant, the authors do note that public universities are almost twice as likely to have a data repository as private universities (see Table 2). This difference is likely due to university systems that offer a data repository for

the member universities. For example, the University of California (UC) system runs the Merritt data repository, which is directly linked to many of the individual UC campuses' library websites. Given the expense of maintaining a data repository, university-system-wide data repositories represent a good use of resources for the universities that can leverage such connections.

Overall, there is a general trend toward libraries offering more data services with larger research universities leading the way. Based on the percentage of libraries offering data support for the six funding levels examined in this study (see Table 3), and estimating growth from previous studies, the authors expect that data services, having a dedicated data librarian, and providing research data hosting will become standard for all research-intensive academic libraries. This research suggests that with growth in higher institutional research output, greater opportunities for development of library data services will also grow.

Data Policies

This study found that just under half of the universities examined (44%) had a data policy of some sort; two-thirds of which are standalone data policies (29% overall) and one-third are intellectual property policies that cover data (15% overall). These numbers are larger than the 18% existence of university data policy that the DataRes project described by Keralis et al. (2013). DataRes studied the top 197 NSF and NIH university grant awardees, searching Google and individual university websites for the keywords "data management" and "policy". While the study is comparable in size, more policies were located in this study than the DataRes projects and several reasons have likely led to this, including the use of more search terms and broader keywords, the specific inclusion of university IP policies in this study, and an expected growth in research data policy as universities have increasingly focused on research data management in the past three years.

Ignoring IP policies, this study measures a 10% difference from the Keralis et al. (2013) project in the prevalence of standalone university data policies. Assuming at least some of this represents an overall increase in data policy, a question that remains is "Where is this difference is coming from?" While specific universities cannot be isolated to identify growth in this area, the data from this study show that universities with data services and a data librarian are statistically more likely to have a standalone data policy. Therefore, the growth in data policy could reasonably correlate with the observed growth in data services, as universities focus more on research data overall. However, no study currently addresses the direct influence of the creation of data services on creation of policy or vice versa. In seeing these strong correlations, the authors hope to discover some causation in future studies.

The greatest challenge in gathering these policies, as also found by the Bohémier et al. (2011) study, lays in the inconsistency across the universities and the inability of the authors to see policies that exist on internal websites. Though it is probable that other data policies exist for these institutions, the myriad locations (Office of Research, Office of the Provost, Library, Office of Sponsored Research, etc.), variety of terms used, and requirements for authentication prevent the casual seeker from easily discovering the policies for a given institution. As the data policy landscape continues to evolve, the authors hope that some standardization will occur in this area.

Data Policy Content

When data policies were discoverable, policies content included definitions of research data, data ownership, data retention, and terms surrounding the separation of a researcher from the institution. However, only data ownership was found to be in greater than half of the policies. These findings differ from previous studies on data policy. In particular, this study found that a significant number of research universities (15%) use their IP policies to cover research data and these IP policies focus almost solely on ownership as compared with standalone data policies. Previous data policies studies did not specifically examine this source of data policies.

Of most notable difference between the policy content analysis in this research and the DataRes project is that the policies discovered in this present study are more specific. The DataRes project provided little content analysis of the policies and summarized them as being advisory statements as opposed to actual institutional policy. Interestingly, Keralis et al. (2013) mention that a perception from their survey respondents saw data management as a development not fully invested in and that policies were therefore not a priority. The current landscape identified here contradicts this and suggests that universities are attempting to get ahead of data policies.

Standalone data policies appear to more broadly cover different areas of data management, though the focus on retention and researcher separation suggests a significant university concern about legal repercussions rather than on sharing and dissemination of data. This focus differs from that of funding agency data policy. The DataRes project, for example, used a word count analysis to suggest that the NIH data policy focuses primarily on data sharing while the NSF policy focuses on disseminating results (Keralis et al., 2013). If universities are creating data policies in response to funding agency mandates, there is clearly a difference in priorities between the two groups. This difference, along with the continued evolution of funder plans for policy, may create challenges in developing institutional data policies.

For the purposes of this study, a comprehensive textual analysis was not undertaken. However, further analysis by the authors is forthcoming to identify specific ownership practices, highlight commonalities in retention, and to suggest best practices for institutions developing data policy. Further work also needs to be done to compare US institutional data policies to those from British, European, Australian, and Canadian institutions.

Opportunities for Librarians

Data services at libraries have passed the point of novelty and are becoming mainstream, as seen in the results of this study. As these services further develop, librarians have a particular opportunity to inform researchers about the policies at their institutions and, where none exist or where the policies are unclear or deemed damaging to research, to argue for and contribute to their development or modification. The findings suggest that librarians are already doing work in this area, as universities with data services and/or a data librarian are statistically more likely to have standalone policies for research data. By becoming well versed on present institutional policies, librarians providing data services can assist researchers in navigating policies at their institutions to meet the requirements of research funders and journals. This assistance is particularly necessary in a data policy landscape that is still in flux. Better knowledge of these policies in their present forms will also allow for improved advocacy to streamline data sharing and reuse.

Academic libraries and librarians are in a unique position to provide insight and guidance in the development and revision of institutional data policies and services. The development of data repositories, increasing in the scope of library-provided data services, and partnerships with other offices and divisions on campus (Office of Research, IT, Sponsored Programs) can further help support the overall data management efforts on campus. By leading the discussion on institutional and researcher needs, librarians can assist in creating policy as part of the overall data management infrastructure.

CONCLUSION

Data services as offered by the library and data librarians are becoming a standard at major research institutions. However, institutional data policies continue to be difficult to identify and many times provide an additional layer of confusion for researchers. In this present landscape, the trend of library-offered data services and hiring or designation of a data librarian will become typical at major research institutions. Importantly, standalone data policies significantly correlate with the presence of data services and a data librarian. While further research is needed to fully analyze the content of these policies, ownership, retention, and access will remain the primary topics of institutional data policies. Funder,

government, and journal policies continue to emerge and evolve. Institutions are developing or revising policies, providing an important opportunity for libraries and librarians to be at the table in the shaping of these institutional policies and improving researcher awareness and compliance.

REFERENCES

Adler, P. (2015, February 20). *AHRQ, NASA, USDA release plans for public access to funded research*. Association of Research Libraries®. Retrieved February 27, 2015, from http://www.arl.org/news/community-updates/3532-ahrq-nasa-usda-release-plans-for-public-access-to-funded-research#.VPB6u_nF9ic

American Geophysical Union. (2013, December). *AGU publications data policy*. Retrieved October 21, 2014, from http://publications.agu.org/author-resource-center/publication-policies/data-policy/

Antell, K., Foote, J. B., Turner, J., & Shults, B. (2014). Dealing with data: Science librarians' participation in data management at Association of Research Libraries institutions. *College & Research Libraries*, *75*(4), 557–574. http://dx.doi.org/10.5860/crl.75.4.557

Bill & Melinda Gates Foundation. (n.d.). *Bill & Melinda Gates Foundation open access policy*. Retrieved January 5, 2015, from http://www.gatesfoundation.org/how-we-work/general-information/open-access-policy

Bohémier, K. A., Atwood, T., Kuehn, A., & Qin, J. (2011). A content analysis of institutional data policies. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (pp. 409–410). New York, NY, USA: ACM. http://dx.doi.org/10.1145/1998076.1998159

Borgman, C. L. (2007). *Scholarship in the digital age information, infrastructure, and the Internet*. Cambridge, Mass.: MIT Press.

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. http://dx.doi.org/10.1002/asi.22634

Bruna, E. M. (2014, September 4). *The opportunity cost of my #OpenScience was 36 hours + $690* (UPDATED). Retrieved from http://brunalab.org/blog/2014/09/04/the-opportunity-cost-of-my-openscience-was-35-hours-690/

Dietrich, D., Adamus, T., Miner, A., & Steinhart, G. (2012). De-mystifying the data management requirements of research funders. *Issues in Science and Technology Librarianship*, *70*(1). http://dx.doi.org/10.5062/F44M92G2

Erway, R., & OCLC Research. (2013). *Starting the conversation university-wide research data management policy*. Dublin, Ohio: OCLC Research. Retrieved from http://www.oclc.org/content/dam/research/publications/library/2013/2013-08.pdf

Evola, M. (2013, July). *Texas Tech University: Office of the Vice President for Research*. Retrieved December 28, 2014, from http://www.depts.ttu.edu/vpr/faculty/scholarly-messenger/stories/2013/July/dataownership.php

Faulkes, Z. (2014, March 7). *NeuroDojo: Who owns data?* Retrieved from http://neurodojo.blogspot.com/2014/03/who-owns-data.html

Fearon, D. J., Gunia, B., Lake, S., Pralle, B. E., & Sallans, A. L. (2013). *Research data management services, SPEC Kit 334 (July 2013)*. Retrieved from http://publications.arl.org/Research-Data-Management-Services-SPEC-Kit-334/

Got data? (2007). *Nature Neuroscience*, *10*(8), 931–931.

Holdren, J. P. (2013). *Increasing access to the results of federally funded scientific research*. Retrieved from http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Information School, University of Sheffield, Cox, A., Verbaan, E., & Sen, B. (2014). A spider, an octopus, or an animal just coming into existence? Designing a curriculum for librarians to support research data management. *Journal of eScience Librarianship*, *3*(1). http://dx.doi.org/10.7191/jeslib.2014.1055

Jahnke, L., Asher, A. D., Keralis, S. D. C., Henry, C., & Council on Library and Information Resources. (2012). *The problem of data*. Washington, DC: Council on Library and Information Resources. Retrieved from http://www.clir.org/pubs/reports/pub154/pub154.pdf

Jahnke, L., & Asher, A. D. (2013). Dilemmas of digital stewardship: Research ethics and the problems of data sharing. In Council on Library and Information Resources, *Research data management: principles, practices, and prospects* (pp. 80–99). Retrieved from http://www.clir.org/pubs/reports/pub160/pub160.pdf

Keralis, S. D., Stark, S., Halbert, M., & Moen, W. E. (2013). Research data management in policy and practice: The DataRes project. *Research Data Management*, 16.

Krier, L., & Strasser, C. A. (2013). *Data management for libraries: A LITA guide*. Chicago: American Library Association.

Kriesberg, A., Frank, R. D., Faniel, I. M., & Yakel, E. (2013). The role of data reuse in the apprenticeship process. *Proceedings of the American Society for Information Science and Technology*, *50*(1), 1–10. http://dx.doi.org/10.1002/meet.14505001051

Law, P., & Act, A. (n.d.). REEXAMINATION OF PATENTS. *Public Law*, *96*, 517.

Lewis, D. (2012, May 24). *Stealing one's own work* [blog]. Retrieved October 21, 2014, from http://nowiknow.com/stealing-ones-own-work/

Marcus, A. (2013a, February 22). Med student loses paper when former boss claims right to data. *Retraction Watch* [blog]. Retrieved October 21, 2014, from http://retractionwatch.com/2013/02/22/med-student-loses-paper-when-former-boss-claims-right-to-data/

Marcus, A. (2013b, October 25). Dispute over data rights forces retraction of obesity paper. *Retraction Watch* [blog]. Retrieved January 4, 2015, from http://retractionwatch.com/2013/10/25/dispute-over-data-rights-forces-retraction-of-obesity-paper/

McGlynn, T. (2014, March 3). *I own my data, until I don't*. Retrieved July 16, 2014, from http://smallpondscience.com/2014/03/03/i-own-my-data-until-i-dont/

Medical Library Association. (2013, June). *JMLA instructions to authors*. Retrieved January 4, 2015, from https://www.mlanet.org/publications/jmla/jmlainfo.html

Mullins, J. L. (2014). The policy and institutional framework. In J. M. Ray (Ed.), *Research data management: Practical strategies for information professionals* (pp. 25–44). West Lafayette, Indiana: Purdue University Press.

Murray-Rust, P. (2010, July 27). *PP3_0.1: Who owns scientific data? Anyone?* Retrieved from http://blogs.ch.cam.ac.uk/pmr/2010/07/27/pp3_01-who-owns-scientific-data/

National Institutes of Health. (2010, October 15). *NIH grants policy statement (10/10) - Part II: Terms and conditions of NIH grant awards, Subpart A: General – File 6 of 6*. Retrieved January 4, 2015, from http://grants.nih.gov/grants/policy/nihgps_2010/nihgps_ch8.htm

National Science Foundation. (2010, November 30). Dissemination and sharing of research results. Retrieved January 4, 2015, from http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Nature Publishing Group. (2014, November 18). Availability of data, material, and methods. Retrieved January 4, 2015, from http://www.nature.com/authors/policies/availability.html#data

Office of Management and Budget. (1999, September 30). *CIRCULAR A-110 REVISED 11/19/93 As further amended 9/30/99*. Retrieved January 4, 2015, from http://www.whitehouse.gov/omb/circulars_a110

Office of Management and Budget. (2014). *e-CFR: Title 2: Grants and agreements, Institution of higher education:, Title 2: Grants and agreements, Institution of higher education: Electronic Code of Federal Regulations § Part 200—Uniform administrative requirements, cost principles, and audit requirements for federal awards.* Retrieved from http://www.ecfr.gov/cgi-bin/text-idx?SID=ed90f54836feb6a994f657188eb05e33&node=2:1.1.2.2.1&rgn=div5

Pham-Kanter, G., Zinner, D. E., & Campbell, E. G. (2014). Codifying collegiality: Recent developments in data sharing policy in the life sciences. *PLoS ONE*, *9*(9), e108451. http://dx.doi.org/10.1371/journal.pone.0108451

Pilat, D., & Fukasaku, Y. (2007). OECD principles and guidelines for access ot research data from public funding. *Data Science Journal*, *6*, OD4–OD11. http://dx.doi.org/10.2481/dsj.6.OD4

Pittman, D. (2012, May 23). *Insubordination in the lab* [blog]. Retrieved October 21, 2014, from http://www.biotechniques.com/news/Insubordination-in-the-Lab/biotechniques-330913.html

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data Is associated with increased citation rate. *PLoS ONE*, *2*(3), e308. http://dx.doi.org/10.1371/journal.pone.0000308

PLOS One. (n.d.). *Sharing of data, materials, and software*. Retrieved January 4, 2015, from http://www.plosone.org/static/policies#sharing

Rohlfing, T., & Poline, J.-B. (2012). Why shared data should not be acknowledged on the author byline. *NeuroImage*, *59*(4), 4189–4195. http://dx.doi.org/10.1016/j.neuroimage.2011.09.080

Schwartz, C. (2012, July 23). *How research into chronic fatigue syndrome turned into an ugly fight*. Retrieved October 21, 2014, from http://www.thedailybeast.com/articles/2012/07/23/how-research-into-chronic-fatigue-syndrome-turned-into-an-ugly-fight.html

Sheridan, M., Quilter, L., Rubinstein, A., Coburn, E., Furfey, J., Walton, J., & Palmer, L. (n.d.). Module 6: Data sharing & reuse policies. In Lamar Soutter Library, University of Massachusetts Medical School (Ed.), *New England collaborative data management curriculum*. Retrieved from http://library.umassmed.edu/necdmc

Smith, M. (2014). Data governance: Where technology and policy collide. In J. M. Ray (Ed.), *Research data management: Practical strategies for information professionals* (pp. 45–59). West Lafayette, Indiana: Purdue University Press.

Strasser, C. (2012, November 20). *Thanks in advance for sharing your data.* Retrieved from http://datapub.cdlib.org/2012/11/20/thanks-in-advance-for-sharing-your-data/

Strasser, C. (2013, January 31). *Collecting journal data policies: JoRD*. Retrieved from http://datapub.cdlib.org/2013/01/31/collecting-journal-data-policies-jord/

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, *6*(6), e21101. http://dx.doi.org/10.1371/journal.pone.0021101

Tenopir, C., Birch, B., & Allard, S. (2012). *Academic libraries and research data services* (p. 54). ACRL. Retrieved from http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf

Mischo, W., Schlembach, M., & O'Donnell, M. (2014). An analysis of data Management plans in University of Illinois National Science Foundation grant proposals. *Journal of eScience Librarianship*, *3*(1). http://dx.doi.org/10.7191/jeslib.2014.1060

University of Manitoba, & Ishida, M. (2014). The New England collaborative data management curriculum pilot at the University of Manitoba: A Canadian experience. *Journal of eScience Librarianship*, *3*(1), 80–85. http://dx.doi.org/10.7191/jeslib.2014.1061

Whitmire, A., Briney, K., Nurnberger, A., Henderson, M., Atwood, T., Janz, M., … Zilinski, L. (2015, April 9). *A table summarizing the Federal public access policies resulting from the US Office of Science and Technology Policy memorandum of February 2013*. Retrieved from http://figshare.com/articles/A_table_summarizing_the_Federal_public_access_policies_resulting_from_the_US_Office_of_Science_and_Technology_Policy_memorandum_of_February_2013/1372041